RAVI VIGNESH

240-898-8518 | Open to Relocation | No immediate sponsorship required (OPT) | ravivigneshravivignesh@gmail.com | LinkedIn

Summary

Software engineer with 3+ years of experience building backend systems, data pipelines, and database-driven applications. Skilled in SQL, Java, C++, Python, TypeScript, AWS, and distributed systems for scalable production services.

Experience

Software Development Engineer Intern May 2025 – August 2025
Amazon Bedrock *Seattle, Washington*

- Improved production incident resolution efficiency by 62% (40→15 min MTTR) by building a distributed backend service using AWS SNS, SQS, and Lambda to automate infrastructure alert remediation across 100+ weekly production incidents.
- Engineered a production RAG system (Bedrock + Kendra) for incident response, reducing engineer context-switching by 67% through instant SOP retrieval from 60+ documents.
- Accelerated incident mitigation by 3× by building Java backend APIs integrated with a React dashboard, improving observability and streamlining operational workflows for production systems.
- Implemented backend parsing and routing logic for CloudWatch alarm data, eliminating 6–8 manual investigation steps per ticket and saving an estimated ~520 engineering hours annually.
- *Tech: Java, AWS (SNS, SQS, Lambda, CloudWatch, Bedrock, Kendra), Distributed Systems, Monitoring, APIs*

Software Engineer Sep 2021 – Jan 2024
NCompass TechStudio

- Built and maintained Java backend services for high-volume financial transaction workflows, processing 5M+ records/day across validation, routing, error handling, internal data access, and downstream system integration.
- Optimized a PostgreSQL-backed validation service using indexing, query-plan tuning, and Redis read-through caching, reducing API P99 latency by 30%.
- Developed an AWS Glue data pipeline moving 10M+ rows/day from RDS/S3 to Snowflake, adding partitioning and data validation checks to improve pipeline runtime by 25%.
- *Tech: Java, Python, SQL, PostgreSQL, Redis, AWS Glue, RDS, S3, Snowflake, CloudWatch, Jenkins CI/CD, REST APIs*

Skills

- **Languages:** Python, Java, C++, JavaScript, TypeScript, SQL
- **Backend:** REST APIs, Microservices, FastAPI, Node.js/Express, Distributed Systems
- **Databases/Data:** PostgreSQL, MySQL, Redis, AWS Glue, RDS, S3, Athena, Snowflake, ETL
- **Cloud/DevOps:** AWS Lambda, SNS, SQS, CloudWatch, IAM, Docker, Kubernetes, Jenkins, GitHub Actions
- **AI/ML:** AWS Bedrock, Kendra, RAG, Embeddings, Vector Search, pandas, NumPy
- **LLM & GenAI:** Prompt Engineering, LLM Agents, Prompt-Driven Debugging
- **Frameworks & Tools:** FastAPI, Node.js/Express, Docker, Kubernetes, React, Git, GitHub Actions, REST APIs

Education

University of Maryland - College Park, MD
Master of Science in Data Science, GPA: 3.9/4.0 *Aug 2024 – May 2026*

Projects

Multi-Tenant API Gateway with Distributed Rate Limiting January 2026

- Designed a Java-based multi-tenant API gateway with per-tenant rate limiting using a Redis sliding-window algorithm, sustaining 12,000+ requests/min in concurrency tests.
- Integrated Resilience4j circuit breakers to isolate downstream failures, reducing recovery time from 45s to 10s in chaos testing.

Semantic Cache & Query Optimization Engine November 2025

- Built a semantic caching layer for LLM APIs (pgvector + Redis) that deduplicates similar prompts before API calls, reducing redundant model invocations by 68%.
- Implemented query rewriting and observability with OpenTelemetry and Prometheus, enabling cache decisioning in 15 ms during benchmarking.

AtmoSense-Seq-Forecast

- Evaluated a Transformer-based air-quality forecasting model over multi-station pollutant time-series data, comparing 48-hour predictions against a persistence baseline.
- Built evaluation scripts and visualizations for MAE/RMSE by horizon, forecast trajectories, and pollutant-level error analysis.